RESEARCH PROPOSAL

A Statistical Study of Machine Learning Techniques

A synopsis submitted to

SWAMI RAMANAND TEERTH MARATHWADA UNIVERSITY, NANDED (M.S.)

For the degree of

Doctor of Philosophy

In

Statistics

Under the faculty of Science and Technology

By

Mr. Digambar Bhausaheb Uphade
M.Sc. (Statistics), SET
Phone No:-7588013127
Email ID: dbuphade@gmail.com

Guided By,

Dr. Aniket Avinash Muley
Assistant Professor & I/C Head,
Department of Statistics,
School of Mathematical Sciences,
Swami Ramanand Teerth Marathwada University,
Nanded-431606. (M.S.) INDIA
Phone No:-9420265128

July 2019

Table of Contents

Sr. No.	Contents	Page No.
1	Abstract	01
2	Introduction	01
3	Review of Literature and Development in the subject	02
4	Objectives of the Research	04
5	Methodology	05
6	Importance of Study	05
7	Proposed Work Plan	06
8	References	07

Abstract:

Machine learning is one of the fastest growing areas in business, industry and research sector. Also, it is widely used in scientific applications such as bioinformatics, medicine, and astronomy [10, 13]. The term machine learning refers to the automated detection of meaningful patterns in data and it is a part of artificial intelligence that helps in developing models. We are surrounded by a machine learning based technology: search engines learn how to bring us the best results, anti-spam software learns to filter our email messages, and credit card transactions are secured by software that learns how to detect frauds [1]. Digital cameras learn to detect faces and intelligent personal assistance applications on smart-phones learn to recognize voice commands. Cars are equipped with accident prevention systems that are built using machine learning algorithms.

We need to discover efficient solution and potentially useful information from a large amount of data and use it in efficient manner. The synopsis aims to understand the important concepts related to machine learning techniques and to solve the problems by using it.

1. Introduction

Now a dayôs very rapid development of machine learning and its use has been expanded to various fields to better classify, analyze, learn, make decisions, or make predictions. The huge amount of data spread and created over the Internet, organizations, and institutes in the field of Medical Science, Business, Agriculture, Education, and E-commerce etc [4, 6, 8, 9, 11]. The explosive rate of data increment is growing thus finding useful information from large amount of data is more difficult. Statistical techniques play an important role to handle these types of datasets to find efficient solutions scientifically, so that we can take decision for further movement [2, 13].

Machine learning techniques are used to operate on large amounts of data to discover hidden patterns and relationships. Bennaceur and Meinke [2] studied Machine Learning (ML) methods for automatically inferring models from data. Machine learning has been successfully applied in many areas of software engineering including: behaviour extraction, testing and bug fixing. Therefore, a better fundamental understanding of ML methods, their assumptions and guarantees can help to identify and adopt appropriate ML technology for new applications. Bennaceur and Meinke [2] presented an introductory survey of ML applications in software engineering, classified in terms of the models they produce and the learning methods they use. They argued that the optimal choice of an ML method for a particular application should be guided by the type of models one seeks to infer. Also, describes some important principles of ML, give an overview of some key methods, and presented examples of areas of software engineering benefiting from ML. Further, they have discussed the open challenges for reaching the full potential of ML for software engineering and how ML can benefit from software engineering methods. However, applications of ML to software analysis are an ongoing and emerging field, where much remains to be done.

2. Review of Literature and Development in the subject.

Castro et al.[3] have applied Machine learning techniques, such as: artificial neural networks (ANN), support vector machines (SVMs), decision trees, and K-nearest neighbor(KNN) algorithms for classification problems for images of fruit. According to their results, the classification of Cape gooseberry fruits by their ripeness level was sensitive to both the color space and the classification technique used. Nanarkar and Chavan [10] conducted survey and evaluate the various classification techniques used for identifying the type of arrhythmia. Their proposed system classified the type of arrhythmia with high accuracy by using a proper combination of preprocessing and classification techniques. The system has focused on supervised learning technique (SVM) and ANN for classification. Sensitivity, Specificity and Accuracy has been used as a metric for performance evaluation. Singh and Agrawal [14] performed a survey covering the strategies and techniques of sentiment classification and demanding situations appeared within the area. The result of their study reveal that sentiment classifiers are seriously dependent on domains or subjects and also found that exclusive varieties of features and classification algorithms are mixed in a green way so as to conquer their man or woman drawbacks and advantage from each other deserves, and finally enhances the sentiment classification overall performance.

Liu et al. [8] employed an artificial neural network (ANN) model to check the relationship between yield and the factors influencing yield. A feedï forward, completely connected, backï propagation ANN was designed to approximate the nonlinear yield function relating corn yield to factors influencing yield. The influencing factors included soil factors, weather factors, and management factors. Random factors were excluded. The ANN parameters tested included network type, network topology, learning rate, initial weights, type of transfer function, and number of training epochs. A portion of the data set, selected by stratified sampling, was withheld from training and used to verify the accuracy of the yield predictions. Sudduth et al. [15] studied the spatial relationship between crop yields and soil and site parameters using several methods and yield maps estimated by projection pursuit regression and neural network analysis. Safa et al. [12] was applied new artificial neural networks to predict rain-fed wheat yield using meteorological data a few days to few months before harvesting.

McQueen et al [9] describes a range of machine learning strategies to problems in agriculture and horticulture. Based on software workbench for experimenting with a variety of techniques on real-world data sets, and describe a cases study of dairy herd management in which culling rules were inferred from a medium-sized database through the different algorithms available in the WEKA workbench, and compare the resulting rule sets. Yeh and Lien [16] aimed at the case of customers default payments in Taiwan and compares the predictive accuracy of probability of default. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. Authors presented the novel Sorting Smoothing Method to estimate the real probability of default. It is observed that, artificial neural network is the only one that can accurately estimate the real probability of default.

Gonzalez-Sanchez et al. [5] highlighted Machine learning (ML) approach to achieve practical and effective solutions for the problem. Further, they compared the predictive accuracy of ML and linear regression techniques for crop yield prediction in

ten crop datasets. Multiple linear regression, M5-Prime regression trees, perceptron multilayer neural networks, support vector regression and k-nearest neighbor methods were ranked. Real data of an irrigation zone of Mexico were used for building the models. Since M5-Prime achieves the largest number of crop yield models with the lowest errors, it is a very suitable tool for massive crop yield prediction in agricultural planning.

Ruß [11] deals with appropriate regression techniques and evaluates four different techniques on selected agriculture data. The results of his study shows that support vector regression can serve as a better reference model for yield prediction. It is computationally less demanding, at least as understandable as the hitherto multi-layer perceptron and, most importantly, produces better yield predictions. Abu Nimeh et al. [1] compared the predictive accuracy of several machine learning methods including Logistic Regression (LR), Classification and Regression Trees (CART), Bayesian Additive Regression Trees (BART), Support Vector Machines (SVM), Random Forests (RF), and Neural Networks (NN) for predicting phishing emails. They investigated the predictive accuracy of six classifiers on a phishing data set. The results showed that, when legitimate and phishing emails are weighted equally, RF outperforms all other classifiers followed by CART, LR, BART, SVM, and NN respectively.

Knudby et al. [7] combined a set of statistical and machine learning models with habitat variables derived from IKONOS data to produce spatially explicit predictions of the species richness, biomass, and diversity of the fish community around two reefs in Zanzibar. Also, estimated live coral cover, structural complexity and habitat diversity, and explored the importance of habitat variables, at a range of spatial scales, in the predictive models using a permutation-based technique. Their findings indicate that structural complexity at a fine spatial scale has the most important habitat variable in predictive models of fish species richness and diversity, whereas other variables such as depth, habitat diversity, and structural complexity at coarser spatial scales contribute to predictions of biomass. Jaikla et al. [6] aimed to develop rice yield prediction procedure using the Support Vector Regression method (SVR), one of the most widely used techniques in data prediction. The prediction method is divided into 3 phases, i.e., soil nitrogen prediction, rice stem weight prediction and rice grain weight prediction. They have compared the results with the commercial software, i.e., DSSAT4 program implementing Crop Simulation Model (CSM-Rice simulation model). The results indicate that our method is comparable with that of the CSMRice simulation model and the error from their model in the acceptable range.

Fortin et al. [4] Deterministic potato growth models hardly rely on driving seasonal field variables that directly characterize spatial variation of plant growth. They evaluated multiple linear regression (MLR) and neural networks (NN) as predictive models of potato yield. Their model showed potential for application in precision agriculture by accounting for temporal and spatial real-time climatic and crop data. Sarkar et al. [13] was presented foundations of Machine Learning before taking a deep dive into Machine Learning pipelines and solving real-world problems. The need for Machine Learning in todayôs world is introduced with a focus on making data-driven decisions at scale. Also, elaborate the various programming paradigms and how Machine Learning has disrupted the traditional programming paradigm. They were covered basic foundational concepts in areas like mathematics, statistics, computer science, Data

Science, data mining, artificial intelligence, natural language processing, and Deep Learning since all of them tie back to Machine Learning and using tools, techniques, methodologies, and processes in various fields. Concepts relevant to the various Machine Learning methods have also been covered including supervised, unsupervised, semi-supervised, and reinforcement learning. Other classifications of Machine Learning methods were depicted, like batch versus online based learning methods and online versus instance based learning methods. A detailed depiction of the CRISP-DM process model was explained to give an overview of the industry standard process for data mining projects. Analogies were drawn from this model to build Machine Learning pipelines, where they focus on both supervised and unsupervised learning pipelines. They brought everything covered together in solving a small real-world problem of predicting grant recommendations for students and building a sample Machine Learning pipeline from scratch. Last but not the least, challenges, and real-world applications of Machine Learning will gives a good idea on the vast scope of Machine Learning and make you aware of the caveats and pitfalls associated with Machine Learning problems.

Sarkar et al. [13] discussed Challenges in Machine Learning which is a rapidly evolving, fast-paced, and exciting field with a lot of prospect, opportunity, and scope. However it comes with its own set of challenges, due to the complex nature of Machine Learning methods, its dependency on data, and not being one of the more traditional computing paradigms. The following points cover some of the main challenges in Machine Learning:

1. Data quality issues lead to problems, especially with regard to data processing and feature extraction.
2. Data acquisition, extraction, and retrieval is an extremely tedious and time Consuming process.
3. Lack of good quality and sufficient training data in many scenarios.
4. Formulating business problems clearly with well-defined goals and objectives.
5. Feature extraction and engineering, especially hand-crafting features, is one of the most difficult yet important tasks in Machine Learning. Deep Learning seems to have gained some advantage in this area recently.
6. Over fitting or under fitting models can lead to the model learning poor representations and relationships from the training data leading to detrimental performance.
7. The curse of dimensionality: too many features can be a real hindrance.
8. Complex models can be difficult to deploy in the real world.

4. Objectives of the Research

Some of the proposed objectives of the research are:

- Analytical study of different optimization techniques used in machine learning.
- To construct prediction models using various machine learning techniques.
- To develop algorithms and statistical models for existing open source data set.
- To validation of the developed model for various kind of datasets.
- To find application of machine learning techniques in crop yield optimization.
- To study the best optimization machine learning technique and their applications in cyber risk prediction through social media sites.

- To perform comparative study of most common predictive techniques used in machine learning.

During the period of research many new objectives may be developed which will be beneficial for the society.

5. Methodology:

In this study, the research can be carried out by analysis of existing machine learning predictive techniques on different datasets. The research requires various different kinds of databases, tools and techniques, machine learning algorithms, scientific languages, different software and packages.

Mainly Machine Learning techniques are classified into following three types:

- **Supervised Machine Learning**

In supervised learning the model is trained with labeled data then the algorithms predict the output from input data. The model becomes in-charge of data and we tell the model what to predict. Classification, regression, k-nearest neighbors (KNN), Naive Bayes (NB), Support Vector machines (SVM) and Decision Trees (DT) algorithms are come under Supervised Machine Learning. [2, 13]

- **Unsupervised Machine Learning**

In unsupervised learning the model is trained with unlabelled data then the algorithms learn to inherent patterns from the input data. In this the model has no idea about the data but tries to find grouping in data. Clustering and Association algorithms come under this type of Machine Learning. [2, 13]

- **Reinforcement Machine Learning**

It involves the model to learn from its past errors and correct it in the next iteration. In this way model improves its results and increase efficiency. [2, 13]

➢ **Software and Packages**

The most used packages for Data Mining

- Python
- R
- SPSS/ SPSS Clementine
- MINITAB
- WEKA
- MATLAB

The research methodology uses above mentioned different software packages to analyze and compare the predictive statistical techniques on different datasets.

6. **Importance of Study/Society application:**

▪ **National and International Status and some diversified applications of Machine learning**:

Some researcher performed foundation and theoretical research work done in some diversified areas. Also, some authors performed real world applications of Machine Learning in the globe and is summarized as follows:

- Product recommendations in online shopping platforms [1]
- Sentiment and emotion analysis [14]

- Medical[10]
- Agriculture[3,4,5,9,11,12,15]
- Fraud detection and prevention [1,16]
- Content recommendation (news, music, movies, and so on) [14]
- Weather forecasting [12]
- Stock market forecasting Market basket analysis [15]
- Customer segmentation [16]
- Object and scene recognition in images and video [3]
- Speech recognition [14]
- E-mail spam filtering [1]

- **Significance of the study**

Sarkar et al. [13] discussed Challenges in Machine Learning which is a rapidly evolving, fast-paced, and exciting field with a lot of prospect, opportunity, and scope. However it comes with its own set of challenges, due to the complex nature of Machine Learning methods, its dependency on data, and not being one of the more traditional computing paradigms. The following points cover some of the main challenges in Machine Learning:

1. Data quality issues lead to problems, especially with regard to data processing and feature extraction.
2. Data acquisition, extraction, and retrieval is an extremely tedious and time consuming process.
3. Lack of good quality and sufficient training data in many scenarios.
4. Formulating business problems clearly with well-defined goals and objectives.
5. Feature extraction and engineering, especially hand-crafting features, is one of the most difficult yet important tasks in Machine Learning. Deep Learning seems to have gained some advantage in this area recently.
6. Overfitting or underfitting models can lead to the model learning poor representations and relationships from the training data leading to detrimental performance.
7. The curse of dimensionality: too many features can be a real hindrance.
8. Complex models can be difficult to deploy in the real world.

7. Proposed work Plan/ Formulation and Structure of Study:

- **Year wise Plan of work and targets to be achieved**

➢ **First Year**

1. To collect and study of review of literature of machine learning techniques.
2. Collection and Study of different machine learning data sets.
3. Study of different machine learning Software packages.

➢ **Second Year**

1. Try to develop statistical model for machine learning.
2. Selection of one of the standard datasets for application of machine learning Predictive Statistical techniques for comparison of results obtained.
3. Detailed study and collaborative analysis of the results obtained.
4. The comparative study of the use of different machine learning Software packages for the same data sets.

- **Third Year**

1. The publication of the findings of the research.
2. Publication of papers in various national and international journals & conferences.
3. Thesis writing.

8. References:

1. Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007, October). A comparison of machine learning techniques for phishing detection. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit* (pp. 60-69). ACM.
2. Bennaceur, A., & Meinke, K. (2018). Machine learning for software analysis: Models, methods, and applications. In *Machine Learning for Dynamic Software Analysis: Potentials and Limits* (pp. 3-49). Springer, Cham.
3. Castro, W., Oblitas, J., De-La-Torre, M., Cotrina, C., Bazán, K., & Avila-George, H. (2019). Classification of Cape gooseberry fruit according to its level of ripeness using machine learning techniques and different color spaces. *IEEE Access*, *7*, 27389-27400.
4. Fortin, J. G., Anctil, F., Parent, L. É., & Bolinder, M. A. (2011). Site-specific early season potato yield forecast by neural network in Eastern Canada. *Precision agriculture*, *12*(6), 905-923.
5. González Sánchez, A., Frausto Solís, J., & Ojeda Bustamante, W. (2014). Predictive ability of machine learning methods for massive crop yield prediction.
6. Jaikla, R., Auephanwiriyakul, S., & Jintrawet, A. (2008, May). Rice yield prediction using a support vector regression method. In *2008 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology* (Vol. 1, pp. 29-32). IEEE.
7. Knudby, A., LeDrew, E., & Brenning, A. (2010). Predictive mapping of reef fish species richness, diversity and biomass in Zanzibar using IKONOS imagery and machine-learning techniques. *Remote Sensing of Environment*, *114*(6), 1230-1241.
8. Liu, J., Goering, C. E., & Tian, L. (2001). A neural network for setting target corn yields. *Transactions of the ASAE*, *44*(3), 705.
9. McQueen, R. J., Garner, S. R., Nevill-Manning, C. G., & Witten, I. H. (1995). Applying machine learning to agricultural data. *Computers and electronics in agriculture*, *12*(4), 275-293.
10. Nanarkar H., & Chawan P.(2018). A Survey on Classification and identification of Arrhythmia using Machine Learning techniques. *IRJET*, 5(2), 446-449.
11. Ruß, G. (2009, July). Data mining of agricultural yield data: A comparison of regression models. In *Industrial Conference on Data Mining* (pp. 24-37). Springer, Berlin, Heidelberg.
12. Safa, B., Khalili, A., Teshnehlab, M., & Liaghat, A. (2004, January). Artificial neural networks application to predict wheat yield using climatic data. In *Proceedings of 20th International Conference on IIPS* (pp. 1-39). Iranian Meteorological Organization.
13. Sarkar, D., Bali, R., & Sharma, T. (2018). Machine learning basics. In *Practical Machine Learning with Python* (pp. 3-65). Apress, Berkeley, CA.

14. Singh, K. P., & Agrawal, S. Sentiment Classification using Machine Learning: A Survey. *International Journal on Recent and Innovation Trends in Computing and Communication*, *5*(5), 212-216.
15. Sudduth, K. A., Drummond, S. T., Birrell, S. J., & Kitchen, N. R. (1996). Analysis of spatial factors influencing crop yield. *Precision Agriculture*, (precisionagricu3), 129-139.
16. Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2), 2473-2480.

Signature
Research Scholar
Mr. Digambar Bhausaheb Uphade

Signature
Research Guide
Dr. Aniket A Muley

Place: - Nanded
Date: - 23/07/2019